

Preface

Bioinformatics is an emerging field in which statistical and computational techniques are used extensively to analyze and interpret biological data obtained from high-throughput genomic technologies. Genomic technologies allow us to monitor thousands of biological processes going on inside living organisms in one snapshot, and are rapidly growing as driving forces of research, particularly in the genetics, biomedical, biotechnology, and pharmaceutical industries.

The success of genome technologies and related techniques, however, heavily depends on correct statistical analyses of genomic data. Through statistical analyses and the graphical displays of genomic data, genomic experiments allow biologists to assimilate and explore the data in a natural and intuitive manner. The storage, retrieval, interpretation, and integration of large volumes of data generated by genomic technologies demand increasing dependence on sophisticated computer and statistical inference techniques. New statistical tools have been developed to make inferences from the genomic data obtained through genomic studies in a more meaningful way.

This textbook is of an interdisciplinary nature, and material presented here can be covered in a one- or two-semester course. It is written to give a solid base in statistics while emphasizing applications in genomics. It is my sincere attempt to integrate different fields to understand the high-throughput biological data and describe various statistical techniques to analyze data. In this textbook, new methods based on Bayesian techniques, MCMC methods, likelihood functions, design of experiments, and nonparametric methods, along with traditional methods, are discussed. Insights into some useful software such as BAMarray, ORIGIN, and SAM are provided.

Chapter 1 provides some basic knowledge in biology. Microarrays are a very useful and powerful technique available now. Chapter 2 provides some knowledge of microarray technology and a description of current problems in using this technology. Foundations of probability and basic statistics, assuming that the reader is not familiar with statistical and probabilistic concepts,

are provided in Chapter 3. Chapter 4 provides knowledge of some of the fundamental probability distributions and their importance in bioinformatics. Chapter 5 is about the inferential process used frequently in bioinformatics, and several tests and estimation procedures are described in that context. Some of the nonparametric methods such as the Wilcoxon and Mann-Whitney tests are useful in bioinformatics and are discussed in Chapter 6. Details of several well-known Bayesian procedures are provided in Chapter 7. Recently, Markov Chain Monte Carlo methods have been used effectively in bioinformatics, and Chapter 8 gives details of MCMC methods in a bioinformatics context. Chapter 9 is about analysis of variance methods that provide an important framework for the analysis of variance. Chapter 10 is about various statistical designs, including modern and classical designs found useful in bioinformatics. Chapter 11 discusses multiple comparison methods and error rates in a bioinformatics context. R-code for examples provided in the text is shown along with respective examples in each chapter. Real data are used wherever possible. Although this textbook is not a comprehensive account of techniques used in bioinformatics, it is my humble attempt to provide some knowledge to the audience in the area of bioinformatics.

AIMS AND OBJECTIVES OF THIS TEXTBOOK

This textbook covers theoretical concepts of statistics and provides a sound framework for the application of statistical methods in genomics. It assumes an elementary level of statistical knowledge and a two-semester sequence of calculus. Most of the textbooks at this level assume an advanced level statistical course and advanced level calculus course, which creates problems for nonstatistics majors and graduate students from nonstatistics disciplines who would like to take a course in statistical bioinformatics to gain knowledge in the interdisciplinary field. This textbook tries to provide a remedy for that problem. Statistics itself is an interdisciplinary applied subject; therefore, many real-life data examples and data are provided to explain the statistical methods in the genomics field. Codes for computer language R are provided for almost every example used in the textbook. Due to the tremendous amount of data, knowledge of using computers for data analysis is highly desirable.

One of the main aims of this textbook is to introduce concepts of statistical bioinformatics to advanced undergraduate students and beginning graduate students in the area of statistics, mathematics, biology, computer science, and other related areas, with emphasis on interdisciplinary applications and use of the programming language R in bioinformatics. This textbook is targeted to senior regular full-time undergraduate level and beginning graduate level students from statistics, mathematics, engineering, biology, computer science, operation research, pharmacy, and so forth. This textbook also is targeted to

researchers who would like to understand different methods useful in understanding and analyzing high-throughput data. This textbook is also useful for people in the industry as a reference book on statistical bioinformatics, providing a solid theoretical base in statistics and a clear approach of how to apply statistical methods for analyzing genomic data, and the use of the computer programming language R in the implementation of the proposed procedures.

The prerequisite for this course is a semester of statistics at the elementary level; however, a higher-level course in statistics is desirable. The material in this textbook can be covered in two semesters. In this textbook, an attempt is made to provide theoretical knowledge to students and develop students' expertise to use their knowledge in real-world situations. At the same time, this textbook provides knowledge for using computer-programming skills in R to implement the methodology provided here.

In this textbook, concepts are developed from ground zero level to the advanced level. If desired, some chapters can be skipped from a teaching point of view, and the remaining material can be covered in one semester. The following flowchart provides options on how to use this textbook for teaching a one-semester course. For a two-semester course, coverage of the complete textbook is recommended.

This textbook is the outcome of classroom teaching of the senior undergraduate and beginning graduate level course in statistical bioinformatics for several semesters, seminars, workshops, and research in the area of statistical bioinformatics. Material in this textbook has been tested in classroom teaching of a one-semester course for several years. The audience was senior undergraduate students and beginning graduate students from many disciplines who had two semesters of calculus, and most of them had an elementary level statistical background.

Codes for R programming provided in this textbook are available on the publisher's website under the link for this textbook. R programs provided in the textbook are basic in nature, and are included just to give an idea about the use of R. Users can improve these programs and make them more elegant and sophisticated. Exercises and solution manuals are available on the publisher's website.

KEY FEATURES

- Bioinformatics is an interdisciplinary field; therefore, the audience is from different fields and different backgrounds. It is hard to assume that the audience will have a good background in statistics or biology. Therefore, this textbook presents the concepts from ground zero level and builds it to an advanced level.